第三章

一元线性回归

Outline

一元线性回归模型可以表示为：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

其中，

- Y_i 被称为因变量（dependent variable）、被解释变量（explained variable）或响应变量（response variable）
- X_i 被称为自变量（independent variable）、解释变量（explanatory variable）或回归元（regressor）
- ε_i 被称为误差项（error term）或扰动项（disturbance），它包含了除 X_i 之外所有其他可能影响 Y_i 的因素。
- β_0 和 β_1 被称为回归系数（coefficients）或模型参数（parameters），其中，β_0 被称为常数项（constant）或截距（intercept），β_1 被称为斜率（slope）

一元线性回归模型

- 假设 3.1：（线性假设）

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (i = 1, 2, \ldots, n).$$

- 假设 3.2：（随机样本）

$\{(X_i, Y_i) : i = 1, 2, \ldots, n\}$ 是从总体中抽取的一个样本量为 n 的随机样本。 (2)

- 假设 3.3：（解释变量存在样本变异）

解释变量 X 的样本观测值 $\{X_i : i = 1, 2, \ldots, n\}$ 不取某一固定的常数。 (3)

- 假设 3.4：（解释变量外生）

$$\mathbb{E}(\varepsilon_i \mid X_i) = 0, \quad (i = 1, 2, \ldots, n). \tag{4}$$

- 假设 3.5：（条件同方差）

$$\mathbb{E}(\varepsilon_i^2 \mid X_i) = \sigma^2 > 0, \quad (i = 1, 2, \ldots, n). \tag{5}$$

β_0 和 β_1 是如下最优化问题的最优解：

$$\min_{b_0,b_1} \mathbb{E}(Y - b_0 - b_1 X)^2.$$

为了估计回归系数，我们可以利用样本均值近似总体期望得到如下最优化问题：

$$\min_{b_0,b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

由于样本量是固定的，上式中的 $1/n$ 可以省略。因此，β_0 和 β_1 的估计量可以通过解如下最优化问题求得：

$$\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2. \tag{6}$$

上式中的目标函数常被称为残差平方和（residual sum of squares），记为 $RSS(\boldsymbol{b})$。通过最小化 $RSS(\boldsymbol{b})$ 得到回归系数估计量的估计方法被称为普通最小二乘（ordinary least squares, OLS）估计。

根据最优化的一阶条件，最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 满足：

$$\frac{\partial}{\partial b_0} RSS(\boldsymbol{b}) = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \tag{7}$$

$$\frac{\partial}{\partial b_1} RSS(\boldsymbol{b}) = -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \tag{8}$$

由式 (7) 可得：

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{9}$$

进一步有：

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2} \equiv \frac{s_{XY}}{s_X^2} \tag{10}$$

普通最小二乘估计实际上是在估计线性投影系数。

根据线性投影的性质，$\mathbb{E}(X_i\varepsilon_i)=0$，$\mathbb{E}(\varepsilon_i)=0$。由矩估计的思想，我们可以利用样本均值代替期望，则线性投影系数估计量（也就是 OLS 估计量）应该满足：

$$\begin{cases} \dfrac{1}{n}\sum_i X_i(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0, \\ \dfrac{1}{n}\sum_i(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0. \end{cases} \tag{11}$$

这恰好就是式 (7) 和 (8) 的一阶条件。

若将 1 看成是特殊的解释变量，常数项 β_0 为其系数，并记 $\boldsymbol{X}_i=(1,X_i)'$，则 ε_i 满足的两个矩条件可以合并为 $\mathbb{E}(\boldsymbol{X}_i\varepsilon_i)=\boldsymbol{0}$，称为 ε_i 与 $\boldsymbol{X}_i$ 正交。

定义回归的拟合值（fitted value）为 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$，残差（residual）为 $e_i = Y_i - \hat{Y}_i$。式 (11) 可以表示为：

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0},$$

或，等价地，

$$\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0}. \tag{12}$$

此式被称为正则方程（normal equation），它表明：在样本中，残差与解释变量正交。因此，OLS 估计的核心是正交条件：在总体中，误差项和解释变量在 $\mathbb{E}(\boldsymbol{X}_i \varepsilon_i) = \mathbf{0}$ 的意义上彼此正交，其样本形式 $\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0}$ 说明 OLS 估计量满足残差与解释变量正交。

$\hat{\beta}_0 + \hat{\beta}_1 X_i$ 通常被称为样本回归线（sample regression line, SRL），它是对总体回归线的估计。

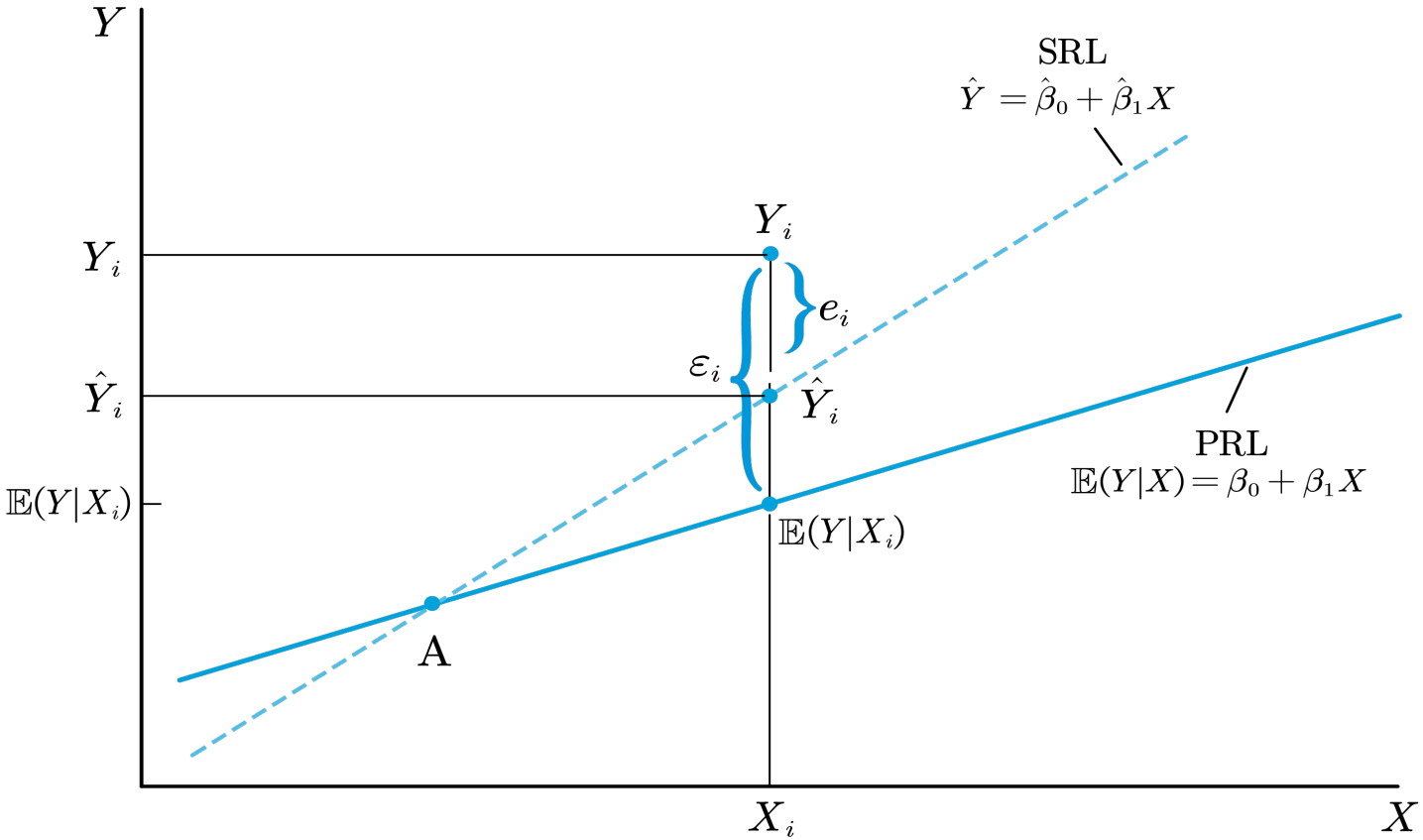

图: 样本回归线和总体回归线

R^2

（中心化）R^2，也称为可决系数（coefficient of determination），衡量的是除常数项之外的解释变量对 Y 的解释能力，其定义如下：

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{13}$$

如果线性回归中含有常数项，则有：

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2. \tag{14}$$

由式 (14)，R^2 可以表示成：

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{15}$$

根据上式，R^2 是由 X 解释了的 Y 的样本变异占其总样本变异的比例。

具有很低的 R^2 的线性回归模型中的系数估计量依然可能是解释变量对被解释变量因果效应的合理估计。

关于 R^2 的更多讨论读者可以阅读一篇知乎上的文章“为什么计量经济学家不看 R-square”，网址为：
https://zhuanlan.zhihu.com/p/19931167。

OLS 估计量的无偏性

若假设 3.1-3.4 成立，则 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的无偏估计量，即

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \quad \mathbb{E}(\hat{\beta}_1) = \beta_1.$$

OLS 估计量的方差（同方差情况）

若假设 3.1-3.5 成立，则以 $(X_1, \ldots, X_n)$ 为条件 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 的条件方差分别是：

$$\mathrm{Var}(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \equiv \frac{\sigma^2}{\mathrm{TSS}_X}, \tag{16}$$

$$\mathrm{Var}(\hat{\beta}_0 \mid X_1, \ldots, X_n) = \frac{\sigma^2 n^{-1} \sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{17}$$

误差项方差的无偏估计

令

$$s^2 \equiv \frac{\sum_i e_i^2}{n-2} = \frac{\text{RSS}}{n-2}.$$

若假设 3.1-3.5 成立，则：$\mathbb{E}(s^2) = \sigma^2$。

若假设 3.4 成立，β_1 为线性投影系数，当 X 为虚拟变量时，可以证明：

$$\beta_1 = \mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0). \tag{18}$$

β_1 可以按照如下方式估计：

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0, \tag{19}$$

其中，$\bar{Y}_1$ 是样本中 $X_i = 1$ 的个体的 Y_i 的样本均值；$\bar{Y}_0$ 是样本中 $X_i = 0$ 的个体的 Y_i 的样本均值。

$\hat{\beta}_1$ 称为 DIM 估计量。

DIM 估计是否可以用来衡量 X 对 Y 的因果效应？这里只分析二值虚拟变量和被解释变量之间的因果关系。

- 这种情况下的因果效应常被称为 X 对 Y 的处置效应（treatment effect）
- 为了强调虚拟变量 X 的特殊性，我们将使用字母 D 来代替 X。
- 虚拟变量处置效应分析的重要应用领域是政策评估（policy evaluation），此时，$D = 1$ 的个体是政策实施的对象，属于处理组（treatment group）；$D = 0$ 的个体不受政策影响，属于控制组（control group）。
- 评估政策的效果就是对处置效应进行估计和检验。

反事实比较在推断因果过程中起到了关键的作用，定义因果效应的主流方法是引入潜在结果（potential outcome）框架。

个体 i 在处置变量 D_i 下的潜在结果定义为：

$$\text{潜在结果} = \begin{cases} Y_{1i} & \text{如果 } D_i = 1, \\ Y_{0i} & \text{如果 } D_i = 0. \end{cases}$$

有了潜在结果的定义，处置 D_i 对个体 i 的处置效应可以容易地定义为：

$$\tau_i = Y_{1i} - Y_{0i}. \tag{20}$$

在实证分析时，研究者通常关注的是平均处置效应（average treatment effect, ATE），其定义为：

$$\tau_{ate} = \mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i} - Y_{0i}]. \tag{21}$$

此外，研究者还可能关注处理组的平均处置效应（average treatment effect on the treated, ATET）和控制组的平均处置效应（average treatment effect on the untreated, ATENT），它们的定义分别为：

$$\tau_{atet} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1], \tag{22}$$

$$\tau_{atent} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]. \tag{23}$$

由于处置效应分析的是 D_i 对 Y_i 的影响，那么一元线性回归是否可以帮助我们推断因果关系？

考虑如下回归方程：

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$

我们知道：

$$\beta_1 = \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0).$$

并且，最小二乘估计量 $\hat{\beta}_1$ 通过 DIM 的方式估计 β_1。于是，问题就转化为 β_1 是否具有处置效应的解释，如果 β_1 是 ATE、ATET 或 ATENT 中的某一种平均处置效应，那么 OLS 就可以利用样本数据估计这种平均处置效应。

以 ATET 为例，我们有：

$$\begin{aligned}
\beta_1 &= \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0) \\
&= \mathbb{E}(Y_{1i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0) \\
&= \mathbb{E}(Y_{1i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 1) + \mathbb{E}(Y_{0i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0) \\
&= \tau_{atet} + \mathbb{E}(Y_{0i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0).
\end{aligned}$$

因此，我们发现 β_1 并不等于 τ_{atet}，而是在 ATET 的基础上加上了 $\mathbb{E}(Y_{0i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0)$，这两个期望的差代表了样本选择偏误 (selection bias)。

既然一元线性回归通常不能直接用来推断因果关系，我们自然会问：若要一元线性回归可以用来推断因果关系，它需要满足什么条件呢？

为了回答这个问题，我们需要建立观测结果和潜在结果之间的关系，即潜在结果模型（potential outcome model, POM）：

$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i}). \tag{24}$$

我们先分析一种简单的情况：对任意个体 i，$Y_{1i} - Y_{0i} = \tau$。此时，我们有：

$$Y_i = Y_{0i} + \tau D_i. \tag{25}$$

现将上式转换为一元线性回归的形式，为此，定义 $\beta_0 = \mathbb{E}(Y_{0i})$，我们有：

$$Y_i = \beta_0 + \tau D_i + \varepsilon_{0i},$$

其中，$\varepsilon_{0i} = Y_{0i} - \mathbb{E}(Y_{0i}) = Y_{0i} - \beta_0$，$\tau$ 为 τ_{ate}。显然假设 3.1 和 3.2 是可以满足的。同时，只要样本里既有处理组个体又有控制组个体，假设 3.3 也是可以满足的。根据无偏性，只要 $\mathbb{E}(\varepsilon_{0i} \mid D_i) = 0$，$\tau$ 的 OLS 估计量 $\hat{\tau}$ 就是 ATE 的无偏估计。而 $\mathbb{E}(\varepsilon_{0i} \mid D_i) = 0$ 等价于 $\mathbb{E}(Y_{0i} \mid D_i) = \mathbb{E}(Y_{0i})$，即 Y_{0i} 均值独立于 D_i。

上述分析假设了个体处置效应 τ_i 为常数 τ，对于更一般的情况，我们有：

$$\tau_i = Y_{1i} - Y_{0i} = \mathbb{E}(Y_{1i}) + \varepsilon_{1i} - [\mathbb{E}(Y_{0i}) + \varepsilon_{0i}] = \tau_{ate} + [\varepsilon_{1i} - \varepsilon_{0i}],$$

其中，$\varepsilon_{1i} = Y_{1i} - \mathbb{E}(Y_{1i})$。由式 (24)，容易得到：

$$Y_i = Y_{0i} + D_i\left[\tau_{ate} + (\varepsilon_{1i} - \varepsilon_{0i})\right] = \beta_0 + \tau_{ate} \times D_i + \varepsilon_i,$$

其中，$\varepsilon_i = D_i(\varepsilon_{1i} - \varepsilon_{0i}) + \varepsilon_{0i}$。为了保证 OLS 估计量无偏，我们需要 $\mathbb{E}(\varepsilon_i \mid D_i) = 0$，即

$$\mathbb{E}(\varepsilon_i \mid D_i) = \mathbb{E}\left[D_i(\varepsilon_{1i} - \varepsilon_{0i}) + \varepsilon_{0i} \mid D_i\right] = 0.$$

使上式成立的充分条件是 $\mathbb{E}(\varepsilon_{1i} \mid D_i) = 0$ 且 $\mathbb{E}(\varepsilon_{0i} \mid D_i) = 0$，也就是 Y_{0i} 和 Y_{1i} 均值独立于 D_i。

Y_{0i} 和 Y_{1i} 均值独立于 D_i 的理想情况是随机化实验。事实上，在随机化实验中，处置状态的分配与潜在结果之间是相互独立的。